ESTADÍSTICA ESPAÑOLA
Vol. 46, Núm. 157, 2004, págs. 409 a 430

Minería de datos y lógica difusa. Una aplicación al estudio de la rentabilidad económica de las empresas agroalimentarias en Andalucía

por
BÁRBARA DÍAZ DIEZ
ANTONIO MORILLAS RAYA
Departamento de Estadística y Econometría
Universidad de Málaga

RESUMEN

En este trabajo se estudia la rentabilidad de la empresa agroalimentaria andaluza mediante un conjunto de ratios, elaborados por el Instituto de Estadística de Andalucía a partir de la Central de Balances de Actividades Empresariales de Andalucía. El objeto es encontrar las características contables de las empresas más rentables. Los aspectos metodológicos que se contemplan en la aplicación comprenden algunas técnicas estadísticas avanzadas y nuevos métodos de extracción de conocimiento en grandes bases de datos (*knowledge discovery y data mining*).

Las conclusiones a que se llega, expresadas en forma de reglas difusas obtenidas de la base de datos mediante la "teoría computacional de la percepción" (Zadeh, 2001; Last, Klein y Kandel, 2001), parecen plenamente congruentes con los postulados del análisis financiero. Si la rotación de activos es baja, no hay altas rentabilidades. Por el contrario, una importante rotación del activo, acompañada por

una aceptable situación de liquidez, es lo que caracteriza a las empresas más rentables.

Palabras clave: Minería de datos, Extracción de conocimiento, Redes Neuronales Artificiales, Árboles de decisión, Ratios contables, Rentabilidad económica.

Clasificación A.M.S. : 94A17, 04A72, 62M45

1. INTRODUCCIÓN

La Central de Balances de Actividades Empresariales de Andalucía se ha elaborado a partir de una muestra de empresas que han presentado sus estados contables en los Registros Mercantiles de Andalucía. La información base versa sobre los Balances de situación, las Cuentas de Pérdidas y Ganancias y las Cuentas de Pérdidas y Ganancias analíticas. Los datos proporcionados por el IEA(1) se refieren, exclusivamente, al ejercicio de 1998 y cubren a 40.601 empresas, clasificadas por provincia y rama de actividad(2), según la CNAE de cuatro dígitos(3). Sobre ellas se han observado 21 variables, en forma de ratios contables. No ha sido posible disponer de ratios referentes a otro ejercicio, lo que hubiera posibilitado algún tratamiento para evitar la marcada influencia coyuntural que puede ir asociada a toda información que se refiera a un único año.

Las 21 variables contempladas y un resumen de los 37.022 casos finalmente contemplados pueden verse en la tabla 1.

(1) Central de Balances de Actividades Empresariales de Andalucía (IEA, 1998). Agradecemos desde aquí la colaboración prestada por este organismo. Para más información, puede verse, también, IEA: *Central de Balances de Actividades Empresariales de Andalucía. Notas Metodológicas*.

(2) Hay 156 empresas no clasificadas por su actividad que no serán contempladas en este estudio.

(3) Por razones de secreto estadístico, se trabajará con una agregación a dos dígitos, realizada con el GESCLA-97 del INE, basada en el Sistema Europeo de Cuentas (SEC, 1995) y adaptada a la clasificación A60 contemplada en el *Sistema de Cuentas Económicas de Andalucía. Marco Input-Output 1995*, del Instituto de Estadística de Andalucía (1999).

Tabla 1

VARIABLES Y RESUMEN DE CASOS

Ratios contables	*Válidos*	*Media*	*Desv. típica*	*Perdidos*	*%*
Generación VAB	37022	-308,3317	59454,5884	0	,0
Generación EBE	37009	43,2662	7490,2301	13	,0
Rotación circulante	35446	2447,7401	221363,4141	1576	4,3
Margen beneficio	37022	-411,3753	66752,1570	0	,0
Rotación recursos propios	37018	27,9223	1930,7523	4	,0
Rentabilidad financiera	37018	-20,8403	9672,4179	4	,0
Coste recursos ajenos	36839	5,9215	152,4593	183	,5
Rotación activo total	37019	4,1409	285,8177	3	,0
Estructura activo	35464	3710,5647	334715,6257	1558	4,2
Autonomía financiera	37018	-19,7351	4822,8238	4	,0
Coeficiente endeudamiento	37017	10,8951	719,9874	5	,0
Tesorería	36634	5,1468	184,7459	388	1,0
Liquidez	36634	4,0343	153,8281	388	1,0
Solvencia	36634	7,2964	249,8427	388	1,0
Garantía	36839	11,0230	360,6369	183	,5
Margen bruto explotación	37022	-341,5102	59502,7017	0	,0
Margen neto explotación	37022	-359,8753	66074,7787	0	,0
Rentabilidad activo fijo	37019	4,7047	742,8491	3	,0
Rentabilidad económica	37019	10,6255	174,2168	3	,0
Rotación activo fijo	35464	279621,8177	23089142,1044	1558	4,2
Disponibilidad	36634	2,1115	130,3071	388	1,0

Fuente: Elaboración propia, a partir de la Central de Balances (IEA).

La inspección directa de esta tabla pone en evidencia un par de cuestiones que deben ser analizadas con precaución:

1. Hay tres variables que presentan un buen número de valores perdidos. En realidad, no es tan importante la cantidad de información que puede ser desaprovechada, con la subsiguiente pérdida de eficiencia, como la posibilidad de que los valores perdidos sigan un patrón no aleatorio(4). En tal caso, tanto ignorarlos como

(4) Véase, Rubin (1976), Little y Rubin (1987) o Schafer (1997), entre otros.

estimarlos, mediante alguno de los sistemas de imputación al uso(5), podrían introducir sesgo o modificar la matriz de covarianzas, respectivamente. Esto sería altamente pernicioso para la obtención de los patrones de ratios contables mediante componentes principales.

2. Los valores medios (por algunos signos y tamaños) y la dispersión (por ciertos valores desmesurados y por la disparidad que presenta entre los distintos ratios) alertan de posibles deficiencias en la base de datos consultada.

Por otro lado, ha sido motivo de una amplia literatura la peculiar problemática de las ratios contables y de su tratamiento estadístico(6):

1. Sobre su elaboración y forma de la distribución, se advierte de cuestiones relacionadas con el cálculo de fracciones, abundancia de valores anómalos por contaminación (diversidad empresarial y sectorial) o derivados de errores de observación, manipulación y de la aleatoriedad, presencia de correlaciones espurias, por definiciones similares de los ratios, y fuerte desviación de la normalidad (con gran asimetría positiva).

2. Dada la diversidad de propuestas de ratios existente para medir un mismo aspecto de la actividad empresarial y las fuertes correlaciones que, por tanto, se observan, es usual asociarlos en grupos o patrones que, obviando la redundancia en la información, expresen, independientemente (única forma de que se puedan interpretar individualmente), las distintas características contables de las empresas. La solución que generalmente se adopta es identificar esos *patrones de ratios contables*, nombre con el que suele conocerse en la literatura, mediante un análisis de componentes principales, unas veces como objetivo final, por su interés intrínseco, otras como paso intermedio para análisis posteriores. En nuestro caso, para el estudio de las características contables de las empresas más rentables, será necesario tomar como variables de entrada los patrones de ratios contables que las definen. Así, pues, la fiabilidad de las componentes principales que se tendrán que obtener es crucial en la aplicación.

Dada la panorámica descrita anteriormente, se hace imprescindible, por un lado, un análisis exploratorio profundo de la base de datos y el empleo de métodos robustos, que hagan que dichas componentes sean menos sensibles a la amplia casuística estadística que se acaba de comentar. Por otro lado, es aconsejable huir

(5) Véase, Demster, Laird y Rubin (1977). El método EM de estos autores es inaplicable en nuestro caso (no normalidad).

(6) Un buen resumen de los temas teóricos y empíricos tratados en la literatura sobre ratios contables puede verse en Salmi y Martikainen (1994), en Aguiar (1989) o en González Pérez (1997). Véase, también, Lev y Sander (1989), Fredka y y Hopwood (1983), Foster (1986), Deakin (1976) y Lev (1978), entre muchos otros.

de técnicas estadísticas clásicas de clasificación, dependientes, generalmente, de hipótesis teóricas, y sustituirlas por el empleo de procedimientos de clasificación de modelo libre, como los que se verán más adelante.

2. ANÁLISIS EXPLORATORIO. PROBLEMAS ESTADÍSTICOS

En el análisis exploratorio se ha comprobado la presencia de todos los problemas comentados anteriormente. Todos las pruebas realizadas sobre el patrón de datos perdidos rechazan la hipótesis de aleatoriedad, resultando, por tanto no ignorable dicho patrón (Rubin, 1976). Esta evidencia, junto a otra serie de circunstancias (datos sobre media y dispersión muy anómalos y correlaciones altas(7) entre sí y con otros ratios), nos ha decidido ha eliminar las tres variables afectadas del análisis.

Además, se han aplicado una serie de filtros lógicos para depurar la información y se han realizado contrastes normalidad, de linealidad en las relaciones entre variables, análisis de heterocedasticidad (sectores) y estudio de valores atípicos (*outliers*), resultando evidente un claro incumplimiento de todas estas hipótesis, como augura la literatura sobre ratios contables.

Aunque en menor medida, a pesar de la depuración efectuada con la ayuda de los análisis exploratorios, los problemas continuaban siendo graves, con incumplimiento de las hipótesis básicas y presencia de valores anómalos por doquier. Por estas razones, se ha procedido a la estimación robusta multivariante de los parámetros de localización, escala y covariación, utilizando el estimador Elipsoide de Mínimo Volumen(8) (MVE). Con los resultados de esta estimación, se ha optimizado el proceso de detección multivariante de los valores anómalos(9) y se ha obtenido una matriz de correlaciones robusta, a partir de la cuál se realiza, finalmente, la extracción de las componentes principales, que son las que pueden observarse en la tabla 2.

(7) Producto de definiciones próximas, lo que puede provocar una estimación inflada de la matriz de covarianzas (Tabachnick y Fidell, 1996).

(8) Rousseeuw y Leroy (1987), Rousseeuw y van Zomeren (1990), Rousseeuw y Hubert (1997), Rousseeuw y van Driessen (1999) y García Pérez (2001).

(9) Una vez efectuado el análisis exploratorio, las medidas obtenidas para los diferentes ratios son absolutamente razonables y totalmente congruentes con otras fuentes (véase, por ejemplo, ESECA, 2001).

Tabla 2

COMPONENTES PRINCIPALES

componente	*Contenido*	*% varianza*	*% varianza ac.*
Liquidez	Componente principal con mayor peso de las ratios de liquidez, tesorería, solvencia, disponibilidad y garantía	33,782	33,782
Márgenes	Componente principal con mayor peso de las ratios margen de beneficio, margen neto de explotación, margen bruto de explotación, rentabilidad de activo fijo, generación EBE.	27,400	61,182
Endeudamiento	Componente principal con mayor peso de las ratios de coeficiente de endeudamiento y rotación de recursos propios	13,170	74,352
Rotaciones	Rotaciones de activo total	11,000	85,353

Las cuatro componentes explican algo más del 85% de la variabilidad incorporada en las variables originales (ratios). Será con estas nuevas variables con las que vamos a investigar las características contables que diferencian las empresas agroalimentarias más y menos rentables de Andalucía. Se tratará, por tanto, de una aproximación tipo *extremos polares*, propia de ciertos análisis discriminantes (véase, Hair y otros, 1998 y Green, Tull y Albaum, 1988), pero que se ha ampliado en este trabajo, incorporando al ámbito del análisis de la rentabilidad con ratios contables la idea de frontera gruesa (*thick frontiere*), procedente de los estudios sobre eficiencia propios de la economía financiera. Así se obtendrá una mejor discriminación entre las empresas más eficientes (rentables) y las menos, a la vez que se aísla un más que posible efecto industria (sector), que es una de las ventajas manifiestas de este tipo de enfoque(10)

(10) Berger and Humphrey, 1991, 1992; Bauer y otros, (1993). El empleo de la frontera gruesa se traduce operativamente en una estratificación de la muestra por sectores, tomando para cada sector aquellas empresas que se encuentran por debajo del primer cuartil y por encima del último en cada uno de ellos. De esta forma, se evita el sesgo que pudiera producirse por las diferencias sectoriales en rentabilidad.

3. EXTRACCIÓN DE CONOCIMIENTO A PARTIR DE LA BASE DE DATOS

La extracción de conocimiento a partir de bases de datos, conocida, también, como minería de datos, se ha definido como "la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de datos," (Frawley and Piatetsky-Shapiro and Matheus, 1992). Se considera a menudo una "mixtura entre estadística, inteligencia artificial e investigación de bases de datos" (Pregibon, 1997). Las primeras obras de referencia en este campo son Piatetsky-Shapiro y Frawley (1991) y Fayyad, Piatetsky-Shapiro, Smyth y Uthurusamy (1996). Otras técnicas utilizadas son los árboles de decisión, la inducción de reglas, las redes neuronales(11), los algoritmos genéticos o la lógica difusa(12).

Los sistemas de inferencia difusos(13) (*Fuzzy Inference Systems - FIS*) pueden ser especialmente interesantes cuando se trabaja con variables lingüísticas o con datos imprecisos. Fue Zadeh (1971) quién sugirió la utilización del concepto de variable lingüística para establecer reglas difusas de tipo IF-THEN en el procesamiento de la información en un sistema de control(14). Para definir las reglas caben dos vías: diseñarlas con base a opiniones de expertos, si no hay información adecuada, o utilizar algún sistema de aprendizaje (red neuronal, por ejemplo) para extraerlas de la base de datos disponible. En este sentido, la definición de los conjuntos difusos que representan los distintos valores de cada variable, es un primer paso en un proceso de razonamiento automatizado en el que se opera sobre la base de percepciones en lugar de sobre medidas de las mismas. Esto es lo que se ha dado en llamar la "teoría computacional de la percepción" (Zadeh, 2001).

(11) La utilización de redes neuronales en el reconocimiento de patrones, está relativamente extendida. La gestión empresarial puede ser uno de sus principales campos de actuación (Ripley, 1994; Altman, Marco y Barreto, 1994).

(12) En Goebel y Gruenwald (1999) puede consultarse una comparativa entre los programas informáticos disponibles sobre minería de datos, atendiendo tanto a su situación legal (comercial, freeware, shareware,..) como a las técnicas que utilizan, tipos y número de variables que permiten utilizar, ... Allí se comenta que la mayoría de ellos utilizan técnicas "estandar" si bien, "las técnicas de otros campos prometedores como los conjuntos difusos o los algoritmos genéticos tan sólo acaban de empezar a encontrar su sitio en el software dedicado a la extracción de conocimiento en bases de datos".

(13) Los más conocidos son el de Mandani y el de Sugeno, cuyo desarrollo formal puede verse en Foulloy y Galichet (1992).

(14) Las reglas difusas combinan uno o más conjuntos difusos de entrada, llamados antecedentes o premisas, y les asocian un conjunto difuso de salida, llamado consecuente o consecuencia. Los conjuntos difusos de la premisa se asocian mediante conjuntivas lógicas del tipo y, o, etc. Una regla típica de tipo IF-THEN sería: Si x es A e y es B, entonces z es C, donde x, y y z son variables lingüísticas y A, B, y C son conjuntos difusos definidos en los universos del discurso X, Y y Z. Los sistemas de control con reglas difusas están ampliamente divulgados y han generado miles de patentes industriales.

En este trabajo se adopta, precisamente, dicho enfoque. Creemos que existe borrosidad en el contexto de aplicación de las componentes principales extraídas. El concepto asociado a dichas componentes es impreciso y se mide de forma aproximada. Llamar LIQUIDEZ, por ejemplo, a una combinación lineal del conjunto de las 20 variables que intervienen en la definición de la correspondiente componente, con muy diferentes pesos, algunas de las cuáles nada tienen que ver con la liquidez, y de la que afirmamos que explica un determinado porcentaje de la varianza total, es sólo una aproximación a la realidad. En cuanto a sus valores, ¿se puede afirmar seriamente que representan fiel y exactamente la intensidad de la liquidez en una determinada empresa?. Nos parece que una variable generada de tal forma adolece de una manifiesta imprecisión. Por tanto, parece aconsejable tratar dicha variable como una variable lingüística, introduciendo en el análisis el razonamiento aproximado, pues precisión y certeza suelen competir en distinta dirección en estos casos (Zadeh, 1973, p. 28). Sus valores (liquidez *baja*, *media y alta*, por ejemplo) o modalidades serán considerados como conjuntos difusos, obtenidos mediante los procedimientos que se exponen a continuación. Posteriormente, estos conjuntos, en forma de números triangulares, se incorporarán al sistema de reglas difusas con el que se obtendrá conocimiento teórico acerca de las características contables de las empresas más rentables.

3.1 Categorización de la variables mediante el algoritmo IFN(15) y reglas asociadas

El método utilizado para agrupar las componentes principales en intervalos y estudiar su asociación (reglas) para definir la rentabilidad se basa en el algoritmo

(15) En opinión de sus autores, se trata de una red neuronal, basada en la teoría de la información. Pero puede discutirse si se trata de una red neuronal o de un árbol de decisión. Según sus autores (véase Last y Kandel, 2001 y Last y Maimon, 2002), la estructura de la red que proponen difiere de la estándar de un árbol de decisión (Quinlan, 1986, 1993) en dos aspectos: 1. La red está restringida al mismo atributo de entrada en todos los nodos de cada capa escondida. 2. Sus conexiones nodo-objetivo representan reglas de asociación entre input y objetivo, mientras que los árboles de decisión estándar se usan sólo para extraer reglas de predicción. Por lo demás, sus autores también ponen de manifiesto que la naturaleza conectiva de su sistema recuerda la estructura topológica de las redes neuronales con capas escondidas. Por otro lado, las redes de información difieren de las neuronales en que los pesos, basados en la información, son definidos solo para las conexiones a la capa objetivo, mientras que las capas internas estás asociadas con valores o intervalos de los atributos de entrada y no tienen peso alguno. Por el contrario, una red neuronal tiene un peso asociado con cada conexión entre capas. Finalmente, señalan que su algoritmo difiere de los más clásicos utilizados en aprendizaje mediante árboles de decisión (CART y C4.5) que utilizan una aproximación *post-pruning* (formando un árbol máximo y podando después), mientras que el suyo realiza una poda previa (*pre-pruning*), deteniendo la construcción de la red cuando ningún atributo causa un descenso significativo de la entropía.

IFN (Info-Fuzzy-Network) de Maimon and Last (2000)(16). Los pesos con que se indica la información contenida en esa asociación o dependencia existente entre las mismas, recogen lo que se llama *información mutua esperada* en la Teoría de la Información(17).

El algoritmo(18) selecciona una variable para entrar en el modelo si contribuye al descenso total en la entropía condicional. En la primera de las capas ocultas, correspondiente a la primera variable de entrada en el sistema, el número de nodos viene determinado por el número de intervalos(19) en que se particione la misma. Estos nodos, que se dividirán solo si proporcionan un descenso estadísticamente significativo en la entropía condicional de la variable de salida(20), se asocian en la siguiente capa a tantas nodos como intervalos haya para la siguiente variable de entrada en el sistema(21). La construcción de la red acaba cuando no existe ninguna variable candidata para entrar al sistema que disminuya significativamente la entropía condicional de la variable de salida. Finalmente, se realiza la conexión entre cada nodo terminal (que no se divide) y cada nodo de la última capa significativa (oculta), con los nodos correspondientes a la variable de salida.

En el caso de nuestra aplicación(22), como puede observarse en la figura 1, los resultados hablan de una red con seis capas, cuatro de ellas ocultas, pertenecien-

(16) Agradecemos a Mark Last, profesor del Departamento "Information Systems Engineering", University of the Negev, (Israel), y creador del programa IFN, sus sugerencias y comentarios, que han permitido la mejora y ampliación de una versión preliminar del trabajo que aquí se presenta.

(17) Véase, Theil (1967; pp. 24-51) y, también, Tilanus y Theil (1965), Gil (1981) y Pardo (1997).

(18) Para ver detalles acerca del modo en que se construye la arquitectura de la red, véase Maimon y Last (2000) , o bien Last, Klein y Kandel (2001).

(19) Si una variable de entrada es continua, se agrupa en intervalos mediante un procedimiento de discretización de la variable, incluido en el algoritmo de construcción de la red, que, de forma reiterada, va buscando los umbrales de partición (reconsiderados cada vez que una nueva variable entra a formar parte del sistema) que maximizan la contribución a la información mutua de la variable de salida

(20) La significación de la disminución en la entropía condicional en un nodo oculto (información mutua entre una variable candidata a input y la variable de salida, dado un nodo oculto) se mide por un test de la razón de verosimilitudes. La hipótesis nula (H_0) es que la variable candidata a input y la variable de salida son independientes, dado un nodo oculto (implicando que su información mutua condicional es cero). Bajo H_0, el estadístico se distribuye como una χ^2 con un número de grados de libertad igual al número de proporciones independientes estimadas a partir de las tuplas asociadas con el nodo oculto.

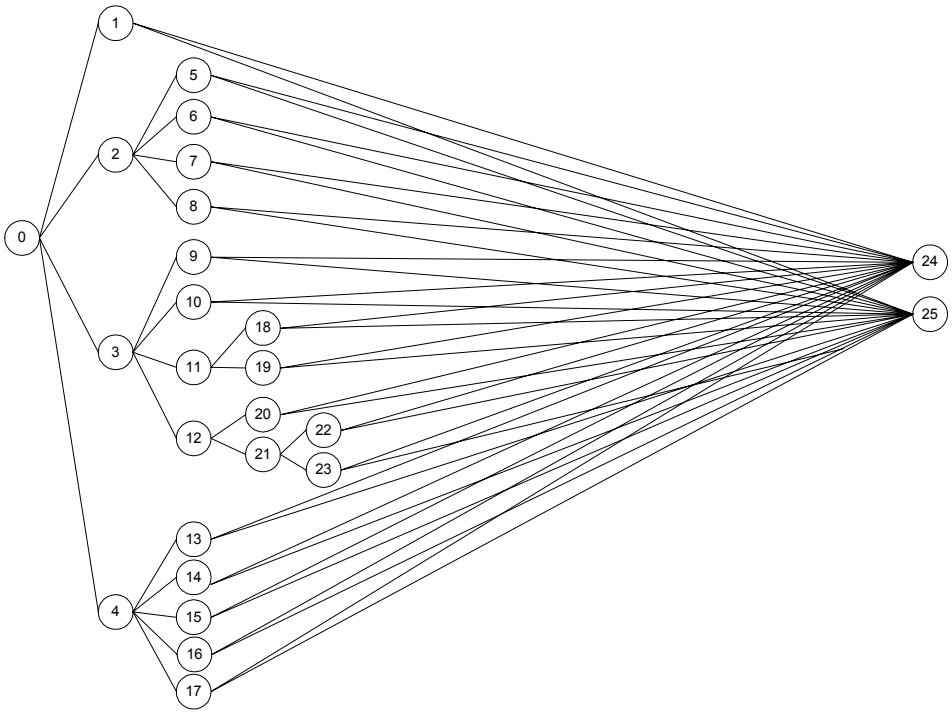
(21) En cualquier caso, un nuevo nodo se crea sólo si hay al menos una tupla asociada a él.

(22) En un primer momento, se han utilizado 1481 casos, correspondientes al primer y tercer cuartil, dejando fuera 59 observaciones (aproximadamente el 5%), seleccionadas de forma aleatoria, para la validación del modelo.

tes a las variables ROTACIONES, LIQUIDEZ, MÁRGENES y ENDEUDAMIENTO. La red consta de 26 nodos, de los que 17 son indivisos y finales. La capa de salida presenta dos alternativas, relacionadas con las empresas más y menos rentables.

Figrua 1

MODELO DE RED PARA LOS PATRONES DE RATIOS



Fuente: Elaboración propia

En la tabla 3 se dan los límites inferiores de los distintos intervalos creados(23) para las variables, así como su código.

(23) Estos intervalos se crean automáticamente mediante el algoritmo descrito.

Tabla 3:

LÍMITES INFERIORES DE LOS INTERVALOS

Variable	*Códigos*				
	0	1	2	3	4
ROTACIONES	-1´8064	-0´8362	-0´5884	-0´1292	
LIQUIDEZ	-2´1014	-0´6631	-0´4467	-0´1936	6´1965
MARGENES	-10´1484	-0´1708			
ENDEUDAMIENTO	-3´2001	1´464			

Entre las dos primeras variables seleccionadas por el modelo, ROTACIONES y LIQUIDEZ, contribuyen en un 95% a la información mutua total (0.708/0.749). Esto significa que estas dos variables son las más determinantes para el valor esperado de la rentabilidad económica de la empresa. La variable ENDEUDAMIENTO es la que menos información aporta, con un descenso en la entropía condicional de la variable de salida del 2´3% (véase la tabla 4).

Tabla 4

ORDEN DE SELECCIÓN DE VARIABLES DE ENTRADA

Iteración	*Variable*	*Entropía condicional*	*I.M. Condicional*	*Información mutua (acumulada)*
0	Rotaciones	0.677	0.323	0.323
1	Liquidez	0.292	0.385	0.708
2	Márgenes	0.257	0.035	0.743
3	Endeudamiento	0.251	0.006	0.749

Una vez que se ha formado la estructura de la red, cada conexión entre un nodo terminal y un nodo de la capa de salida representa una regla probabilística entre una conjunción de valores de las variables de entrada y valores de salida. Un nodo z representa una conjunción de valores de variables de entrada. Posteriormente, se da un peso(24) a cada regla de asociación entre un nodo terminal (que no se divide) y un valor de salida V_j .

En nuestro caso, se han obtenido un total de 28 reglas con peso asociado distinto de cero. Dichas reglas proceden de la combinación de 17 nodos terminales (que no se dividen) y finales (de la última capa oculta) a los dos nodos de la capa de salida. En la tabla 5 se exponen las reglas con mayores pesos asociados obtenidas con el modelo.

Tabla 5

REGLAS CON MAYOR PESO

Regla 18. Si ROTACIONES es mayor que –0´1292 y LIQUIDEZ está entre –0´1936 y 6´1965 entonces ROA es alta (Peso= 0´2707)
Regla 2. Si ROTACIONES está entre –1´8064 y –0´8362 entonces ROA es baja (Peso =0´1974)
Regla 15. Si ROTACIONES es mayor que –0´1292 y LIQUIDEZ está entre –2´1014 y –0´6631 entonces ROA es baja (Peso =0´0833)
Regla 27. Si ROTACIONES está entre –0´5884 y –0´1292 y LIQUIDEZ está entre –0´1936 y 6´1965y MÁRGENES es mayor que –0´1708 y ENDEUDAMIENTO está entre –3´2001 y 1´464 entonces ROA es alta (Peso= 0´0767)
Regla 16. Si ROTACIONES es mayor que –0´1292 y LIQUIDEZ está entre –0´4467 y –0´1936 entonces ROA es alta (Peso= 0´0586)
Regla 9. Si ROTACIONES está entre –0´5884 y –0´12921292 y LIQUIDEZ está entre –2´1014 y –0´6631 entonces ROA es baja (Peso =0´0438)

(24) Los pesos de conexión que unen los nodos indivisos y los nodos de la capa final a los nodos de la capa de salida se calculan: $w_z^{ij} = P(V_{ij}; z) \bullet \log \frac{P(V_{ij} / z)}{P(V_{ij})}$, dónde $P(V_{ij}; z)$ es la probabilidad conjunta estimada del valor V_{ij} y el nodo z, $P(V_{ij} / z)$ es la probabilidad condicional (a posteriori) estimada del valor V_{ij}, $P(V_{ij})$ es la probabilidad incondicional (a priori) estimada del valor V_{ij}. Las probabilidades vienen dadas por las frecuencias relativas.

Se observará que no es fácil valorar de forma inmediata los resultados obtenidos. Los valores de los distintos patrones de ratios dados mediante las componentes obtenidas, positivos y negativos, son difícilmente interpretables, ni siquiera en términos comparativos. Es más, carece de sentido pensar, por ejemplo, en rotaciones negativas. La utilización de términos lingüísticos, contextualizando los valores observados mediante conjuntos difusos, permite, sin embargo, como se verá más adelante, la expresión e interpretación de estas reglas de una forma más razonable.

La capacidad predictiva(25) del modelo generado a partir de la formación de estas reglas se ha comprobado con un 5% de los datos elegido aleatoriamente y que no se utilizaron para la modelización(26). En el conjunto de los 1422 casos utilizado para el entrenamiento(27) de la red, se clasificaron correctamente 1320, es decir, un 92´8%. En el conjunto de validación(28), se clasificaron correctamente 57 de los 59 casos, es decir, un 96´6%. Por tanto, las diferencias en rentabilidad parecen estar en el modelo bien caracterizadas por los conceptos recogidos en las componentes principales. Se ha constatado también que el modelo es un buen predictor para observaciones extramuestrales. Es de destacar, además, que el porcentaje de aciertos en los dos grupos (empresas más y menos rentables) está bastante equilibrado, como puede observarse en la tabla 6.

(25) Dados unos valores para cada variable de entrada al sistema, el valor para la variable de salida se obtiene seleccionando un valor j* que maximice la probabilidad condicional (a posteriori) estimada de la variable de salida i en el nodo z $(P(V_{ij} / z)), j^* = \arg\max_j P_j(V_j / z)$, de tal forma que éste será el valor predicho de la variable de salida i en el nodo z.

(26) Para poder llevar a cabo la generalización de los resultados obtenidos es preciso evitar que el modelo aprenda con datos que posteriormente se van a clasificar. Para ello hay que llevar a cabo algún procedimiento de validación que arroje una medida fiable de la capacidad predictiva del modelo, generalmente dada por el porcentaje de aciertos.

(27) La precisión de la predicción en los datos de entrenamiento es una estimación optimista del verdadero valor de predicción. En cualquier caso, la validez del modelo se puede comprobar comparando la precisión de la predicción en los datos de entrenamiento (mediante el ratio de error, en nuestro caso de 0´072) con el ratio de error a priori (error de predicción), si no se utiliza modelo para la predicción excepto la estimación. Dados unos valores para cada variable de entrada al sistema, el valor para la variable de salida se obtiene seleccionando un valor j* que maximice la probabilidad condicional (a posteriori) estimada de la variable de salida i en el nodo z $(P(V_{ij} / z)), j^* = \arg\max_j P_j(V_j / z)$, de tal forma que éste será el valor predicho de la variable de salida i en el nodo z.

(28) En el conjunto de validación, el ratio de error es 0´034.

Tabla 6

MATRIZ DE CLASIFICACIÓN Y VALIDACIÓN

		ROA	*Grupo de pertenencia pronosticado*		*Total*
			1	2	
Original	Recuento	1	664	49	713
		2	53	656	709
	%	1	93´13	6´87	100
		2	7´48	92´52	100
Validación	Recuento	1	26	1	27
		2	1	31	32
	%	1	96´3	3´7	100
		2	3´125	96´875	100

Por si la utilización del 95% de los casos en el grupo de entrenamiento pudiera haber dado lugar a un "sobreaprendizaje", se han llevado a cabo 10 simulaciones en las que sólo se han tomado como tal conjunto el 60% de las observaciones. El 40% restante se ha utilizado para validar la capacidad de predicción. Los resultados obtenidos son satisfactorios, especialmente si se tiene en cuenta que se reduce drásticamente la información disponible a algo menos de 900 observaciones, cosa poco conveniente para este tipo de modelos. El porcentaje de aciertos en entrenamiento oscila entre el 92,29% y el 95,81%. Para el grupo de validación, estos porcentajes son un mínimo del 88,48% y un máximo del 92,52%, estando 7 de las 10 muestras por encima del 90%. Por otra parte, el programa arroja un valor del 93,75% para la estabilidad del sistema en una simulación con 5 conjuntos de validación cruzada con una muestra del 60%, elegida al azar.

3.2 Sistema de inferencia difuso

Dado que las reglas obtenidas se expresan en términos rígidos *(crisp)*, definiendo las fronteras exactas de cada intervalo con valores en nada relacionados con el concepto que intentan representar, su interpretación, credibilidad y poder de representación real quedan seriamente limitados. Téngase en cuenta, además, la imprecisión que, como comentamos anteriormente, arrastran las variables utilizadas. En

situaciones como esta que nos ocupa, puede ser más adecuado extraer conocimiento expresándolo en términos lingüísticos, más cercanos al modo de razonamiento humano y a la forma en que, generalmente, se expresa la teoría sobre un fenómeno económico. Por otra parte, el número de reglas rígidas extraído suele ser muy elevado (un total de 28, en nuestro caso, si exceptuamos las de peso nulo) en comparación a lo que un gestor se plantearía generalmente como esquema para la toma de decisiones(29)

El método que se utiliza a continuación, basado en la teoría computacional de la percepción, permite extraer un conocimiento más claro e inmediato del fenómeno estudiado, relajando la rigidez de los intervalos y reduciendo el número de reglas en tres pasos: en primer lugar, haciendo difusas las reglas; posteriormente, reduciendo el número de las mismas mediante resolución de conflictos; por último, combinando reglas del conjunto reducido(30)

Para hacer difusas las reglas, se proporciona para cada variable de entrada un conjunto de términos lingüísticos: *bajo, medio y alto*(31). En la tabla 7 puede observarse para cada término el número difuso triangular asociado, obtenido bajo los criterios señalados en la nota a pie número 31.

(29) En cualquier caso, en nuestro ejemplo, las seis reglas de mayor peso cubren el 97% de la información mutua, lo que ya supone una reducción considerable del número de reglas.

(30) Se utilizan implicaciones de tipo Mamdani. Si el peso de la regla rígida es positivo, se construye viendo la relación directa entre el antecedente y el consecuente de la regla, mientras que si es negativo, la relación es inversa. El cálculo del grado de pertenencia de cada regla rígida, en el paso a regla difusa se realiza como sigue: $\mu_R = w \cdot \left[\prod_{i=1}^{N} \max_j \{\mu_{Aij}(V_i)\} \right] \cdot \max_k \{\mu_{Tk}(o)\}$. Dónde w es el peso de la regla rígida correspondiente, N es el número de condiciones simples en la regla, V_i es el valor rígido de la condición simple i en la regla rígida (punto medio del intervalo), o es el valor rígido de la salida de la regla (punto medio del intervalo de salida), $\mu_{Aij}(V_i)$ es el grado de pertenencia de la condición simple i al término j y $\mu_{Tk}(o)$ es el grado de pertenencia del valor de salida o en el término k.

(31) Se han tomado como valores de referencia los extremos de las variables, la mediana y primer y tercer cuartil. De esta forma, los conceptos "bajo" , "medio" y "alto" son contexto-dependientes, al tratarse de empresas bajo el primer cuartil, entre el primero y el tercero, y por encima del tercer cuartil.

Tabla 7

TÉRMINOS LINGÜÍSTICOS

Variable	*Término*	*Prototipo*	*Mínimo*	*Máximo*
LIQUIDEZ	Baja	-2.10	-2.10	-0.24
	Media	-0.24	-0.61	0.36
	Alta	9.22	-0.24	9.22
MÁRGENES	Bajos	-10.15	-10.15	-0.10
	Medios	-0.10	-0.47	0.41
	Altos	4.43	-0.10	4.43
ROTACIONES	Bajas	-1.81	-1.81	-0.10
	Medias	-0.10	-0.71	0.47
	Altas	6.73	-0.10	6.73
ENDEUDAMIENTO	Bajo	-3.20	-3.20	-0.11
	Medio	-0.11	-0.58	0.43
	Alto	5.44	-0.11	5.44

Dado que las reglas obtenidas incluyen asociaciones entre conjuntos de valores de entrada (para las distintas variables) y todos los valores posibles de la variable de salida, es posible que varias de ellas puedan tener el mismo antecedente y distinto consecuente. Además, varias reglas pueden diferir en sus valores numéricos, pero ser idénticas en sus términos lingüísticos. De esta forma, las 28 reglas difusas que se obtienen, pueden ser inconsistentes. Por esta razón, para resolver este conflicto, se calcula el grado de cada regla sumando los grados de pertenencia de todas las reglas difusas idénticas y eligiendo de cada grupo en conflicto el valor de salida de la de mayor pertenencia. Por otra parte, reglas con el mismo consecuente pueden unirse mediante el conectivo “o”, es decir, utilizando un operador de unión difuso (generalmente el máximo).

El conjunto reducido de reglas, una vez que se han hecho difusas y se han combinado, puede verse en la tabla 8.

Tabla 8

REGLAS DIFUSAS

Regla 1. Si las ROTACIONES son bajas entonces la RENTABILIDAD es baja (grado:0.1278)
Regla 2. Si las ROTACIONES son medias o altas y la LIQUIDEZ es baja entonces la RENTABILIDAD es baja (Grado=0.0156)
Regla 3. Si las ROTACIONES son altas y la LIQUIDEZ es media o alta entonces la RENTABILIDAD es alta (Grado=0.0036)
Regla 4. Si las ROTACIONES son medias, la LIQUIDEZ es media o alta y los MARGENES son bajos entonces la RENTABILIDAD es baja (Grado=0.0009)
Regla 5. Si las ROTACIONES son medias, la LIQUIDEZ es media y los MARGENES son altos entonces la RENTABILIDAD es alta (Grado=0.0003)
Regla 6. Si las ROTACIONES son medias, la LIQUIDEZ es alta, los MARGENES son altos y el ENDEUDAMIENTO es medio entonces la RENTABILIDAD es alta (Grado=0.0006)

El número de reglas se ha reducido drásticamente y las expresiones lingüísticas con que se definen aportan conocimiento teórico inmediato. Puede observarse que las tres primeras reglas, con diferencia, son las más importantes. Parece claro, según la información que aparece en las reglas, que la componente que mejor caracteriza las diferencias entre las empresas con alta rentabilidad y las menos rentables, sería la que representa las rotaciones de activo. Las empresas con baja rotación de activos se caracterizan, en cualquier caso, por tener baja rentabilidad económica. Por otra parte, por sí sola, la presencia de una rotación de activos media o alta, no garantiza la existencia de alta rentabilidad, ya que, como se observa en la regla dos, cuando la liquidez es baja, las empresas tienen baja rentabilidad. En la regla tres se observa que unas altas rotaciones acompañadas de media o alta liquidez son indicio de alta rentabilidad. Finalmente, según se desprende de las restantes reglas, empresas con rotaciones medias acompañadas de liquidez media o alta, necesitan tener altos márgenes y bajo nivel de endeudamiento para formar parte del grupo de las empresas más rentables.

Analizando la información por sector y provincia, se observa que las rentabilidades medias más altas coinciden con el sector agrario de las provincias de Huelva y Almería. Como se sabe, gran número de este tipo de explotaciones (cultivos forza-

dos) se caracterizan por tener una alta rotación de activo (ingresos de explotación / activo total), lo que vendría a apoyar la validez de los resultados obtenidos.

4. CONCLUSIONES

Como enfoque alternativo al análisis clásico multivariante para el estudio de la rentabilidad empresarial mediante ratios contables, se ha hecho uso de un método que no depende para su aplicación de hipótesis teóricas, que claramente no se dan en la distribución de este tipo de variables contables. La aproximación, tipo extremos polares, mediante la idea de frontera gruesa, permite una más clara caracterización de las empresas con mayor rentabilidad, a la vez que protege de posibles sesgos procedentes de las agrupaciones sectoriales y/o provinciales.

El significado de las componentes extraídas para agrupar ratios situados bajo un mismo patrón y simplificar la dimensión del problema, da como resultado unas nuevas variables (componentes principales) caracterizadas por su vaguedad conceptual y por la imprecisión derivada de la información que incorporan. Su tratamiento como variables difusas, por tanto, nos parece no sólo plenamente justificado, sino, incluso, lo más adecuado.

En las relaciones entre ratios contables, la definición de reglas de comportamiento mediante proposiciones lingüísticas permite una interpretación conceptual de la realidad que no es posible conseguir mediante otras técnicas. Con tan sólo seis reglas difusas (tres básicamente), que expresan directamente conocimiento teórico, se ha condensado la información que permite establecer la discriminación entre empresas de alta o baja rentabilidad económica, haciéndose patente la importancia de tener una alta rotación de las inversiones para alcanzar un alto nivel de rentabilidad. Es ésta una característica propia de los cultivos forzados de Huelva y Almería, que acogen las empresas más rentables, según la base de datos.

Por otra parte, hay que decir que el uso del algoritmo IFN, en conjunción con la teoría computacional de la percepción, ofrece una alternativa realmente interesante a la formación de las reglas a partir de la observación o la intuición, y permite una adecuada “extracción de conocimiento” de una gran cantidad de datos. La red ha captado las relaciones más importantes entre las variables y el sistema de reglas difuso las ha interpretado directamente en forma de conocimiento aproximado.

REFERENCIAS

ALEXANDER, I. Y MORTON, H. (1990): «An Introduction to neural computing». London, Chapman and Hall.

ALTMAN, E. I.; MARCO, G. Y BARRETO, F. (1994): «Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks (the Italian experience)». *Journal of Banking and Finance*, vol. 18, pp. 505-529.

BAUER, P. W.; BERGER, A. N. Y HUMPHREY, D. B. (1993): «Efficiency and productivity growth in U.S. banking», en Fried, H. O.; Lowell, C. A. K. y Schmidt (eds.): *The measurement of productive efficiency: techniques and Applications*, Oxford University Press, pp. 386-413.

BERGER, A. N.(1993): «Distribution free' estimates of efficiency in the U.S. banking industry and tests of the standard distributional assumptions», *The Journal of Productivity Analysis*, 4, pp. 261-292.

BERGER, P. W. Y HUMPHREY D. B. (1991): «The dominance of inefficiencies over scale and product mix economies in banking», *Journal of Monetary Economics*, 20, pp. 501-520.

BERGER, P. W. Y HUMPHREY D. B. (1992): «Measurement and efficiency issues in commercial banking», en Griliches, Z. (ed.): *Ouput measurement in the service sectors*, University of Chicago Press.

DEMPSTER, A.P.; LAIRD, N.M. Y RUBIN, D.B. (1977): «Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)». *Journal of the Royal Statistical Society*, *Series B*, 39, pp. 1-38.

ESECA (2001): *Análisis económico financiero de la empresa andaluza, 2001*. Eseca.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Y UTHURUSAMY, R. (1996): «Advances in knowledge discovery and data mining». Cambridge, MIT Press.

FAYYAD, U. Y IRANI, K. (1993): «Multi-interval discretization of continuous-valued attributes for classification learning». Proceedings 13th International Conference on Artificial Intelligence, San Mateo, CA, pp. 1022-1027.

FOSTER, G. (1986): «Financial statement analysis». New Jersey, Prentice Hall.

FOULLOY, L. Y GALICHET, S. (1992): «Les applications des ensembles flous». Deuxiemes Journees Nationales, pp. 129-136.

FRAWLEY, W., PIATETSKY-SHAPIRO G. Y MATHEUS, C. (1992): «Knowledge Discovery in Databases: An Overview». *AI Magazine*, pp 213-228.

GARCÍA PÉREZ, A. (2001): «Métodos avanzados de estadística aplicada». Madrid, Universidad Nacional de Educación a Distancia.

GIL, P. (1981): «Teoría matemática de la información». Ed. ICE

GOEBEL, M. Y GRUENWALD, L. (1999): «A survey of data mining and knowledge discovery software tools». *SIGKDD Explorations*, vol. 1, nº 1, pp. 20-33.

HASSOUN, M.H. (1995): «Fundamentals of Artificial Neural Networks». Masachusetts, MIT Press.

HAYKIN, S. (1994): «Neural Networks. A comprehensive foundation». New York, Macmillan Publishing Company.

KULLBACK, S. Y LEIBLER R. A. (1951): «On Information and Sufficiency», *Annals of Mathematical Statistics* 22**,** pp. 79-86.

LAST, M. Y KANDEL, A. (2001): «Fuzzification and Reduction of Information - Theoretic Rule Sets» in «Data Mining and Computational Intelligence». Kandel, A., Last, M y Bunke, H (Eds). *Studies in fuzziness and soft computing*, vol 68, pp. 63-93. Physica-Verlag.

LAST, M., KLEIN, A. Y KANDEL, A. (2001): «Knowledge Discovery in Time Series Databases». *IEEE Transactions on Systems, Man and Cybernetics*, vol. 31, Part B, nº 1, pp. 160-169.

LAST,M. Y MAIMON, O. (2002): "«A Compact and Accurate Model for Classification»". *IEEE Transactions on Knowledge and Data Engineering*.

LITTLE, R.J.A. Y RUBIN, D.B. (1987): «Statistical analysis with missing data». New York, John Wiley and Sons.

MAIMON, M. Y LAST, M. (2000): «Knowledge Discovery and Data Mining- The Info-Fuzzy Network (IFN) Methodology». Kluwer Academic Publishers.

PARDO, L. (1997): «Teoría de la información estadística». Ed. Hespérides.

PIATETSKY-SHAPIRO, G. Y FRAWLEY, W. J. (1991): «Knowledge discovery in Databases». MIT Press.

PREGIBON, D.(1997): «Data mining». *Statistical Computing and Grafics*, vol 7,nº 8.

QUINLAN, J.R. (1986): «Induction of Decision Trees». *Machine Learning*, vol 1, nº1, pp. 81-106.

QUINLAN, J.R. (1993): C4.5: «Programs for Machine Learning». Morgan Kaufmann**.**

RIPLEY, B.D. (1994): «Neural networks and related methods for classification». *Journal of the Royal Statistical Society, B*, 56, pp. 409-456.

ROUSSEEU, P. J. Y LEROY, A. M. (1987): «Robust Regression and Outlier Detection». Wiley.

ROUSSEEUW, P. J. Y VAN DRIESSEN, K. (1999): «A fast algorithm for the minimum covariance determinant estimator». *Technometrics*, 41, pp. 212-223.

ROUSSEEUW, P. J. Y HUBERT, M. (1997): «Recent developments in PROGRESS», en Dodge, y.(ed.): *L1-Statistical Procedures and Related Topics*, IMS Lecture Notes, volume 31, pp. 201-214.

ROUSSEEUW, P.J. Y VAN ZOMEREN, B.C. (1990): «Unmasking multivariate outliers and leverage point». *Journal of American Statistical Association*, 85, pp. 633-639.

RUBIN, D. B. (1976): «Inference and missing data». *Biometrika*, 63, pp. 581-592.

SCHAFER, J.L. (1997): «Analysis of incomplete multivariate data». London, Chapman and Hall.

SCHERER, F.M.; ROSS, D. (1990): «Industrial market structure and economic performance». 3th ed., Boston, Houghton Mifflin Company.

TILANUS, C.B. Y THEIL, H. (1965): «The Information approach to the evaluation of input-output forecast». *Econometrica*, vol 32, nº4, pp. 847-862.

TABACHNICK, B.G. Y FIDELL, L.S. (1996): *Using multivariate statistics*. 3ª edición. Harper Collins.

ZADEH, L.A. (1971): «Quantitative fuzzy semantics», *Information Sciences*, 3, pp. 159-176.

ZADEH, L.A. (1973): «Outline of a New Approach to the Analysis of com`plex systems and decision Processes». IEEE Transactions on Systems, Man and Cybernetics, SMC-3,1. pp.28-44.

ZADEH, L.A. (2001): «A New Direction in AI - Toward a Computational Theory of Perceptions», *AAAI Magazine*, pp. 73-84.

DATA MINING AND FUZZY LOGIC. AN APPLICATION TO THE STUDY OF THE PROFIT VALUE OF THE ANDALUSIAN AGRARIAN INDUSTRY

ABSTRACT

In this paper we study the profit value of the Andalusian agrarian industry, from an available assembly of ratios included in the database of the Institute of Statistics of Andalusia Andalusian related to the Head Office of Balances of Business Activities. The aim is to find the accountant characteristics of the most profitable businesses. The methodological aspects we have used in the application are advanced statistical techniques for the exploratory analysis of data, strong estimation or fuzzy neural networks and new knowledge discovery methods and data mining in large databases.

The fuzzy rules extracted from the database by means of the "Computational Theory of Perceptions" (Zadeh, 2001; Last, Kein y Kandel, 2001), are in agreement with financial analysis theory. The businesses with low rotation of assets are characterized, in any case, by having low economic profit value and the combination of high rotations with medium or high liquidity is indicative of high profit value.

Key words: Data mining, knowledge discovery, exploratory data analysis, neural networks, decision trees, accounting ratios, return of assets.

A.M.S classification: 94A17, 04A72, 62M45